Bradter U, Mair L, Jönsson M, Knape J, Singer A, Snäll T.

Can opportunistically-collected Citizen Science data fill a data gap for habitat suitability models of less common species?

***Methods in Ecology and Evolution* 2018**

DOI: https://doi.org/10.1111/2041-210X.13012

Methods in Ecology and Evolution

DR UTE BRADTER (Orcid ID : 0000-0001-5687-1233)

DR JONAS KNAPE (Orcid ID : 0000-0002-8012-5131)

DR ALEXANDER SINGER (Orcid ID : 0000-0002-2777-3789)

Article type : Research Article

Handling editor: Dr Barbara Anderson

Can opportunistically-collected Citizen Science data fill a data gap for habitat suitability models of less common species?

Bradter, U[1*]., Mair, L[1, 2]., Jönsson, M.[1], Knape, J.[3], Singer, A.[1] & Snäll, T[1].

[1]: Swedish Species Information Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden

[2]: School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, UK

[3]: Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

*: Corresponding author: Ute Bradter, Swedish Species Information Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden, Email: ute.bradter@slu.se

Abstract

1. Opportunistically-collected species observations contributed by volunteer reporters are increasingly available for species and regions for which systematically collected data are not available. However, it is unclear if they are suitable to produce reliable habitat suitability models (HSMs), and hence if the species-habitat relationships found and habitat suitability maps produced can be used with confidence to advice conservation management and address basic and applied research questions.

2. We evaluated HSMs with opportunistically-collected observations against HSMs with systematically collected observations. We enhanced the opportunistically-collected presence-only data by adding inferred species absences. To obtain inferred absences, we asked individual reporters about their identification skills and if they reported certain species consistently and combined this information with their observations. We evaluated several HSM methods using a forest bird species, Siberian jay (*Perisoreus infaustus*), in Sweden: logistic regression with inferred absences, two versions of MaxEnt, a model combining presence-absence with presence-only observations and a Bayesian site-occupancy-detection model.

3. All HSM methods produced nationwide habitat suitability maps of Siberian jay that agreed well with systematically collected observations (AUC: 086-0.88) and were very similar to a habitat suitability map produced from the HSM with systematically collected observations (Spearman rho: 0.94-0.98). At finer geographical scales there were differences among methods.

4. At finer scale, the resulting habitat suitability maps from logistic regression with inferred absences agreed better with results from systematically collected-observations than other methods. The species-habitat relationships found with logistic

regression also agreed well with those found from systematically collected data and with prior expectations based on the species ecology.

5. Synthesis and application: For many regions and species, systematically collected data are not available. By using inferred absences from high-quality opportunistically-collected contributions of few very active reporters in logistic regression we obtained HSMs that produced results similar to those from a systematic survey. Adding high-quality inferred absences to opportunistically-collected data is likely possible for many less common species across various organism groups. Well performing HSMs are important to facilitate applications such as spatial conservation planning and prioritization, monitoring of invasive species, understanding species habitat requirements or climate change studies.

Sammanfattning

1. Opportunistiskt rapporterade artobservationer av allmänheten blir alltmer tillgängliga för arter och regioner för vilka systematiskt insamlade data saknas. Det är emellertid oklart om dessa data är användbara som bas för att producera artutbredningsmodeller och därmed om de resulterande artutbredningskartorna tillförlitligt kan användas för naturvårdsprioriteringar och för att besvara grundläggande och tillämpade forskningsfrågor.

2. Vi utvärderade artutbredningsmodeller baserade på opportunistiskt insamlade artobservationer jämfört med modeller baserade på systematiskt insamlade artobservationer. Fokusart var den skogslevande fågeln lavskrika (Perisoreus infaustus) i Sverige. Vi kompletterade opportunistiskt insamlade förekomstdata med icke-förekomstdata. För att erhålla icke-förekomstdata frågade vi först enskilda frivilliga rapportörer om deras förmåga att känna igen fågelarter och om de rapporterade vissa arter konsekvent, och därefter kombinerade vi denna information med deras artobservationer. Vi utvärderade flera

statistiska modelleringsmetoder: logistisk regression med icke-förekomster, två versioner av MaxEnt, en modell som kombinerar en delmängd förekomster och icke-förekomster med observationer av enbart förekomster, och en Bayesiansk modell som tar hänsyn till att rapportören eventuellt inte upptäckte lavskrikor som fanns på en plats.

3. Alla modelleringsmetoder producerade rikstäckande artutbredningskartor för lavskrika som överensstämde väl med systematiskt insamlade observationer (AUC: 086-0.88), och artutbredningskartor baserade på systematiskt insamlade observationer (Spearman rho: 0,94-0,98). Vid finare geografisk upplösning fanns dock skillnader mellan metoder.

4. Vid finare upplösning överensstämde de resulterande artutbredningskartorna baserade på logistisk regression med icke-förekomster bättre med resultat från systematiskt insamlade data än andra metoder. De förklarande miljövariabler som identifierades med logistisk regression överensstämde vidare med variablerna som identifierades utifrån systematiskt insamlade data och med förväntningar baserade på artens ekologi.

5. Syntes och tillämpning: För många arter och regioner är systematiskt insamlade artdata inte tillgängliga. Genom att komplettera opportunistiskt insamlade förekomster med högkvalitativa icke-förekomster från ett fåtal mycket aktiva rapportörer, och sedan använda dessa i logistisk regression, erhöll vi artutbredningsmodeller och -kartor som liknar de från en systematisk undersökning. Det är troligtvis möjligt att komplettera med högkvalitativa icke-förekomstdata för många andra, mindre vanliga arter från olika organismgrupper från frivilligt insamlade data. Tillförlitliga artutbredningsmodeller är viktiga för rumslig naturvårdsplanering och prioritering, övervakning av invasiva arter och förståelsen av arters habitatkrav eller svar på klimatförändringar.

Keywords

Biological records, forest bird, habitat model, niche model, observer behaviour, presence-only, pseudo-absences, species distribution model

Introduction

Habitat suitability models (HSMs) are an important tool in conservation and applied ecology (Franklin 2009). HSMs can be used to produce maps of predicted habitat suitability, which for example are important in spatial conservation planning and prioritization (Elith & Leathwick 2009), monitoring of invasive species (Kadoya *et al.* 2009), mapping ecosystem services (Polce *et al.* 2013) or habitat-climate change studies (Mair *et al.* 2017a). HSMs can also be used to infer species habitat requirements (Franklin 2009). This information is required by conservation practitioners (Braunisch *et al.* 2012) and is useful for basic and applied research (Guisan & Zimmermann 2000).

HSMs use information about a species occurrence at sample locations paired with information about environmental conditions (Franklin 2009). Systematic surveys can provide such species information. Examples are the North American Breeding Bird Survey (www.pwrc.usgs.gov/bbs) or the Swedish Bird Survey (SBS) (www.fageltaxering.lu.se). They are designed to provide representative and comparable data with survey protocols specifying where, when and how to survey.

Systematic surveys are however only implemented in some countries and for some species groups (Isaac *et al.* 2014) leaving large data gaps. Even in countries and for groups covered by systematic surveys, data gaps can remain because systematic surveys often provide few data for rare species, species of localized habitats or species which are active outside of

survey times (e.g. at night) (Bibby, Burgess & Hill 1992). For example, fewer than 25 individuals per year were recorded by the SBS for 41% of 242 species analyzed (Snäll *et al.* 2011).

Naturalists are often particularly interested in the rarer species. They can submit their opportunistic observations, often called Citizen Science (CS) data, to specific databases (Silvertown 2009; Devictor, Whittaker & Beltrame 2010). We refer to opportunistic observations as CS data, in contrast to volunteer-based efforts with systematic sampling designs, like the SBS. CS can provide information even in traditionally data-poor areas, as demonstrated by eBird (http://ebird.org), a global CS database for bird observations (Amano, Lamming & Sutherland 2016). The large amount of data CS can collect is exemplified by the Swedish Species Observation System (www.artportalen.se). Although it only started in 2000, it has now registered > 60 million observations of plants, animals and fungi (www.artportalen.se, accessed 22 Sep 2017) in a country with approximately ten million people.

While large amounts of data can be quickly collected by CS, the lack of a systematic survey design impacts their use (Yoccoz, Nichols & Boulinier 2001; Snäll *et al.* 2011; Isaac *et al.* 2014; Kamp *et al.* 2016). A problem for HSMs can be that CS data are often geographically biased: they contain more data from areas with higher population density and easier access (e.g. through roads) (Mair & Ruete 2016; Tye *et al.* 2017) or from biodiversity hotspots like protected areas (Higa *et al.* 2015; Isaac & Pocock 2015).

For CS data, reliable measures of observer effort are often not available and the information that a species was not observed (“absence”) may not be recorded. A large number of such

opportunistic presence-only records are available from the Global Biodiversity Information Facility (www.gbif.org), the largest online database holding species records of all organism groups from all over the world. Some CS databases, such as ebird, optionally record absence information by providing complete species lists for a region (Sullivan *et al.* 2009). Participants are asked to report all observed species on the list, which creates presence information for the observed and absence information for the unobserved species. However, this is not feasible for all organism groups. For some groups, such as fungi, only some experts can identify all species and the number of complete species lists they could provide would be small. There is also incomplete knowledge of the species occurring for some regions preventing the formulation of complete species lists. Hence, there are CS data with some absence information while others consist of presence-only data.

Even if absences are not collected, it may be possible to infer some by adding information about the consistency of reporting and species identification skills of individual reporters. If reporters can identify a focal species and always report it when seen (i.e. on any visit to any location), a location for which such reporters have submitted observations of other species, but not of the focal species, becomes an inferred absence location (Snäll *et al.* 2011; Mair *et al.* 2017b). Reporting consistency is however rarely taken into account to infer species absence information.

The limitations of CS data can complicate the modelling of species distributions. It is well-known that for MaxEnt, a popular method for HSMs with presence-only data (Phillips, Anderson & Schapire 2006; Elith *et al.* 2011), geographical bias can lead to a predicted habitat suitability map that combines the distribution of the species with the pattern of where reporters go (Elith *et al.* 2011; Yackulic *et al.* 2013). Options to decrease this effect of geographical bias have been suggested (Phillips & Dudík 2008; Phillips *et al.* 2009). Other

HSMs methods, such as logistic regression, are more robust to geographical bias, but require additional absence data (Zadrozny 2004; Phillips *et al.* 2009; Elith *et al.* 2011). Site-occupancy-detection models have performed well under geographical bias (Higa *et al.* 2015). They add an extra component: the probability of detection of the species by an observer (MacKenzie *et al.* 2003; Kéry, Gardner & Monnerat 2010; Kéry & Schaub 2012). Additionally, a recent method combining presence-absence with presence-only data adjusts for geographical bias by estimating it (Fithian *et al.* 2015).

Logistic regression with inferred absences, presence-absence/presence-only and site-occupancy-detection models have rarely been evaluated for HSMs using CS data (but see e.g. Higa *et al.* 2015; Fletcher Jr *et al.* 2016; Mair *et al.* 2017b). Additionally, few studies have assessed HSMs with CS data against independently and systematically collected data (e.g. Phillips *et al.* 2009; Syfert, Smith & Coomes 2013). Due to a lack of systematically collected data for many species and regions, this will only be possible in some cases, but ultimately it is an important validation. CS can collect large amounts of data and potentially be a valuable resource to help address conservation problems in a rapidly changing world. It is therefore important to know whether results from HSMs using CS data are comparable to those using systematically collected data and to understand how well different modelling methods cope with geographical bias.

The aim of our study was to evaluate the performance of several HSM methods using CS data recorded as presence-only against independently and systematically collected SBS data. As a focal species we used the Siberian jay, a forest bird species declining in parts of its European range (Bird Life International 2016). Specifically, we evaluated 1) the predictions from CS models against the observations from the SBS, and 2) the agreement between habitat suitability maps from CS models and from a model with SBS data. 3) We

further evaluated the suggested species-habitat relationships from CS models for consistency with those suggested by the SBS model and for consistency with expectations based on the species ecology. The latter can be interpreted as support for a model (Snäll *et al.* 2014). We built models covering most of Sweden (ca. 1500 km in length).

Materials and methods

Study species

Siberian jay is a forest specialist with a preference for older forest (mature forest henceforth) (Brotons *et al.* 2003; Edenius, Brodin & White 2004; Griesser & Lagerberg 2012). We studied Siberian jay because it: 1) is easily identified, thus being an example of a species for which CS data has a high potential to fill data gaps, 2) is a well-studied species allowing us to assess whether species-habitat relationships suggested by models are realistic, 3) is a less common species, so we expected several of the most active reporters to be strongly motivated to always report it when seen, a pre-requisite for obtaining inferred absences, and 4) has been negatively impacted by modern forest management (Griesser & Lagerberg 2012) whereby our study can have an added benefit in facilitating the conservation of the species.

Systematically collected SBS data

Standard routes of the Swedish Bird Survey (SBS) form a square of 2x2 km and are distributed along a regular grid with 25 km resolution across Sweden (Ottvall *et al.* 2007). A proportion of the routes is surveyed once per year between May and July. Observations were aggregated into the periods 2000–2002, 2003–2007, and 2008–2013, which matches with forest predictor variables available at 5-year intervals (see below). We placed a 2x2 km

square, which approximated the size of Siberian jay territories (observation units henceforth, see below) at the corners of survey squares. Observation units with at least one Siberian jay observation within a period became presences, and absences otherwise (Fig 1a, b). For details see Appendix S1.

Opportunistically-collected CS observations

We used Citizen Science (CS) observations of the focal species from the Swedish Species Observation System for 2000-2013 (Appendix S2). To infer absences we sent a questionnaire to very active forest bird reporters (Appendix S3) and used observations from those that stated they were able to identify the focal species by sight and sound and always reported it when seen. We removed uncertain or wrong observations (Appendix S2), those with large location uncertainties (>500 m) and, to keep CS data independent from SBS test data, observations from the SBS (Appendix S2). We aggregated data into the same year-periods as the SBS data. Locations with at least one observation of the focal species within a period became presences. Locations without observations of the focal species and exceeding criteria for observer effort (> 5 bird species recorded; if locations were very close, the one with higher observer effort was selected; Appendix S2) became absences (Fig. 1e, f). We placed each presence and absence location at the centre of an observation unit.

Environmental data

We constructed environmental predictor variables (Appendix S4) based on existing knowledge of the species ecology (Appendix S4), broadly in three categories: Mature Forest, Forest and Non-forest (Table 1). Collinearity between predictor variables was limited as variable inflation factors were below five, a recommended cut-off value (Zuur *et al.* 2009).

We calculated environmental predictor variables within observation units, unless otherwise stated. Due to the relatively large variation of reported sizes of the year-round territories of Siberian jay (0.4 – 5 km^2) (Edenius, Brodin & White 2004; Nystrand *et al.* 2010; Pukkala *et al.* 2012), we started by comparing CS models with observation units of 1x1 km and 2x2 km using logistic regression (Appendix S5). As the results were very similar between both sizes and the location uncertainties of CS observations would have been large relative to the size of observation units if using the smaller size, we used 2x2 km in the HSMs.

We calculated forest predictor variables for the years 2000, 2005 and 2010 from the nationwide forest raster data in Sweden (25 m resolution). These raster are produced at 5-year intervals, linking Landsat imagery and measurements from the repeat field National Forest Inventory (Reese *et al.* 2003). The extent of the rasters varied in the mountains between years. We included the raster cells that had data in all years. We used raster data for both total forest volume and age to characterize mature forest. Age is likely to be a good indicator for resources such as availability of lichen as storage space for the food hoarding Siberian jay (Cramp & Perrins 1994). Volume is likely a better indicator of denser forest in marginal areas with sparser and more stunted forest growth, such as in the mountains and the north of Sweden. Denser tree growth is important to provide cover for Siberian jay nests (Griesser & Lagerberg 2012; Pukkala *et al.* 2012).

Modelling method for SBS data

We modelled Siberian jay presence-absence (679 presences, 5683 absences, Table S1) using a Binomial distribution and a logit link. Observations in adjacent observation units and in adjacent periods may be from the same individual because Siberian jays often remain in the same territory for life once established (Griesser *et al.* 2007). We accounted for this by

fitting a generalized linear mixed model with survey route as a random effect and the environmental predictor variables as fixed effects (Zuur *et al.* 2009).

Modelling methods for CS data

MaxEnt

MaxEnt is a machine learning algorithm that models species distributions from presence-only data (Phillips, Anderson & Schapire 2006). It can fit very flexible relationships (e.g. quadratic, hinge, threshold) (Phillips, Anderson & Schapire 2006; Phillips & Dudík 2008). As other methods we used do not by default fit such flexible relationships, we disabled all but linear features to facilitate comparisons among methods. Selected quadratic and interaction terms were instead added manually (Appendix S4).

MaxEnt compares environmental information between presence and available locations (background) (Phillips, Anderson & Schapire 2006; Elith *et al.* 2011). To decrease the effect of geographical bias, it has been suggested to draw background data from locations where species impacted by a similar geographical bias as the focal species have been observed (target-group background) (Phillips & Dudík 2008; Phillips *et al.* 2009). We created a target-group-background using 39 bird species, which were seen and reported in similar circumstances as Siberian jay (Appendix S6).

We evaluated two versions of MaxEnt: presences (n = 2865, Table S1) were paired with 10000 background cells selected 1) randomly from a 2x2 km raster placed over the study area (MaxEnt-Random) and 2) randomly from raster cells that were part of the target-group background (MaxEnt-TGB).

Logistic regression

From 60 reporters answering our questionnaire, we identified 38 that stated they always reported Siberian jay when seen and were able to identify the species by sight and sound. From their 2,003,193 observations from 2000-2013 we removed some observations (see above) and aggregated observations from the same location and time period. This resulted in 4758 inferred absences of Siberian jay (Appendix S2, Table S1).

We modelled presence-absence of Siberian jay using a Binomial distribution and a logit link. We used two versions: inferred absences were paired with 1) 2865 presences (Table S1) of all reporters (PresAbs-all) to facilitate a comparison with MaxEnt models using the same presences and 2) 960 presences (Table S1) reported by the same 38 reporters (PresAbs-38) to facilitate a comparison with the presence-absence/presence-only model (see below) using the same presence-absence data.

Presence-absence/presence-only model

The presence-absence/presence-only (PresAbs-PresOnly) model jointly models presence-only data for several species, which are assumed to be affected by the same geographical data collection bias, together with systematically collected presence-absence data (Fithian *et al.* 2015). By estimating the geographical bias, the model for the focal species can be improved. To study the performance of PresAbs-PresOnly when systematically collected data are not available, we used the CS presence-inferred-absence data from the 38 reporters (Table S1). For presence-only data we used presences from all reporters (Table S1). PresAbs-PresOnly models require additional environmental variables to estimate geographical bias in data collection. We used variables representing human population

density and ease of access to locations (Table S2, Appendix S4). For data on the additional species see Appendix S2. PresAbs-PresOnly models also use background data and we randomly selected 40000 cells from a 2x2 km raster placed over the study area.

Site-occupancy-detection model

Site-occupancy-detection (Occ-Det) models estimate the observation process and the ecological process in one model to account for imperfect detection of the species (Kéry, Gardner & Monnerat 2010; Kéry & Schaub 2012). The detection probability is estimated from repeat observations at the same location. For variables used to explain detection probability, see Table S2 and Appendix S7. We modelled the detection probability as the fraction of days during which Siberian jays were recorded out of the number of days the observation unit was visited, per year, using a Binomial distribution and a logit link. We used data from the 38 reporters and modelled the ecological process using a Bernoulli distribution and a logit link.

Variable selection and method evaluation

We used AIC (Akaike's information criterion, Burnham & Anderson 2004) for model selection for all methods except the site-occupancy-detection model (Appendix S8). In other words, we chose the model with the lowest AIC as the best model. In site-occupancy-detection models we dropped variables when their 95% credible interval included zero (Appendix S7). Maps of the predicted probability of Siberian jay occurrence are presented for the SBS model (Fig 1c) and the site-occupancy-detection model (Appendix S7). For each best model we produced a habitat suitability map for the study area and evaluated methods as follows (Appendix S8):

1) We evaluated the predicted habitat suitability values from the best CS model per method against the observations of the SBS using AUC (area under the receiver-operating curve).

2) We calculated the rank-correlation coefficient between the habitat suitability map from the best CS model per method and the habitat suitability map from the best SBS model.
3) We assessed the directions of relationships found in CS models for consistency with relationships found in the SBS model. We also assessed all models for consistency with expectations: based on previous studies on Siberian jay ecology (Appendix S4), we had expectations to find positive relationships to variables in the Mature Forest category, to the percentage of non-mature forest (PercOther), and that a relationship to the percentage of spruce (PercSpruce) would be positive (Table 1).

Software used

The analysis was carried out in R (Version 3.3.3) (R Core Team 2016). MaxEnt models were fit with package dismo (Hijmans *et al.* 2016) and PresAbs-PresOnly models with multispeciesPP (Fithian *et al.* 2015). The site-occupancy-detection model was fitted with JAGS 4.2.0 (Plummer 2003) using Bayesian techniques.

Results

Across Sweden habitat suitability values from all CS methods agreed well with the independently and systematically collected SBS observations (AUC: 086-0.88) (Fig. 2a). The habitat suitability maps from all CS methods (Fig. 1g-l) were also very similar to the habitat suitability map from the SBS model (Fig. 1d) (Spearman rho: 0.94-0.98, Fig. 2b).

The SBS model discriminated well between presences and absences (block-cross-validated AUC: 0.89 +/- 0.03). As expected, the probability of Siberian jay presence increased with predictor variables of the Mature Forest category (PercMature, MeanAge), with elevation,

thought to be a proxy for availability of mature forest at larger spatial scales (Appendix S4) and with non-mature forest (PercOther) (Table 2).

After excluding the easy-to-predict areas outside of the Siberian jay range, maps of all CS methods still showed good agreement to the SBS map (Spearman rho: 0.81-0.96) and habitat suitability scores still agreed well with SBS observations (AUC: 0.73-0.77, Fig. 2). At even finer scales at which management decisions are often taken, agreement with SBS observations and with the SBS habitat suitability map varied strongly regionally and among modelling methods (Fig. 2). For all methods, eastern regions agreed better with the SBS compared to adjacent western regions. Within western regions, the southern regions agreed better with the SBS compared to the northern-most region. This is likely at least partly explained by the strong regional temporal inconsistency in the satellite-derived forest age data, which affected mainly western regions and particularly the north-western region (Appendix S9). At fine scales, the logistic regressions with inferred absences (PresAbs-all, PresAbs-38) agreed overall better and more consistently with the SBS compared to the methods with presence-only data (MaxEnt and PresAbs-PresOnly) or the site-occupancy-detection (Occ-Det) model.

The agreement with the SBS overall decreased when we randomly selected fewer inferred absences (10% - 90% of available absences) in PresAbs-38 (Fig S1). The random sample drawn strongly influenced results, as indicated by the high variability in AUC and Spearman rho values (Fig S1) for different random samples of the same size.

Directions of species-habitat relationships of both logistic regressions (PresAbs-all, PresAbs-38) were largely consistent with relationships in the SBS model for best models (Table 2) and

with expectations based on the species ecology. In contrast, in the presence-absence/presence-only model, MaxEnt and the site-occupancy-detection model, some relationships were of opposite direction compared to the SBS model and compared to expectations (red in Table 2).

The positive effect of mean age of surrounding mature forest patches (Neighbourhood variable), which was found in many models, was due to mean ages much larger than our mature forest threshold value (50 years) (Appendix S10). Suitable forest sites in these heavily managed landscapes are thus more likely to be occupied with increasing ages of surrounding mature forest patches. Residuals of CS models were spatially autocorrelated, but our robustness analysis to residual spatial autocorrelation and to correlation amongst predictor variables showed that neither changed our conclusions (Appendix S11).

Discussion

Global databases collect large numbers of species occurrence records, often without any absence information. We added inferred absences retrospectively, by using information about reporting consistency and species identification skills of reporters. Inferring species absences via a reporter questionnaire was a comparatively small effort likely feasible in many regions. It requires that reporters consistently report the focal species, which is generally more likely for rarer compared to common species, and for species of conservation concern. Logistic regression models from opportunistically-collected presence and inferred absence data produced results that were very similar to those obtained from systematically collected data. This shows the potential of CS data to construct useful habitat suitability models and to facilitate answering basic and applied ecology questions.

Method evaluation

All CS methods produced Sweden-wide habitat suitability rank map patterns that were very similar to the rank pattern from the independently and systematically collected SBS data. Systematically collected data do not necessarily always produce good habitat suitability models. Siberian jay for example use a relatively large area. With only one site visit per year, the SBS data likely contain many false Siberian jay absences. Despite this, the SBS model performed well. Therefore it was reasonable to treat the SBS model as a reference, against which to assess the performance of our CS models.

Results from the logistic regressions with inferred absences agreed overall somewhat better with results from systematically collected surveys compared to methods developed for presence-only data (PresAbs-PresOnly and MaxEnt), and the site-occupancy-detection model. The predicted habitat suitability maps from the logistic regressions had higher and more consistent agreement with the SBS observations and with the map predicted from the SBS observations particularly at the finer scales, which are often relevant for conservation management. The species-habitat relationships from the logistic regressions were also more consistent with the species ecology and the SBS model. This can result in more realistic projections, for example with future land-use scenarios, compared to models which show inconsistencies with a species ecology (Randin *et al.* 2006).

Several studies have recommended the use of site-occupancy-detection models for CS data (Kéry, Gardner & Monnerat 2010; Kéry *et al.* 2010; van Strien, van Swaay & Termaat 2013; Isaac *et al.* 2014; Higa *et al.* 2015) although for bird population trends in Denmark they produced mixed results (Kamp *et al.* 2016). Not taking the detection process into account can bias covariate estimates towards zero (Kéry, Gardner & Monnerat 2010). However, the site-occupancy-detection model for the Siberian jay suggests a low risk for erroneously missing relevant ecological signal due to observation bias. Environmental variables that

strongly influenced Siberian jay occupancy (winter temperature and elevation) did not explain Siberian jay detection probability, and detection probability was not correlated with occurrence probability. Therefore, it is valid to compare the site-occupancy-detection model with the SBS model, which does not take the detection process into account.

The PresAbs-PresOnly model requires a small sample of systematically collected presence-absence data (Fithian *et al.* 2015). As systematic data is scarce for many regions and species groups, we simulated a situation where no systematically collected data, but inferred absences from opportunistically-collected observations were available. With this data the PresAbs-PresOnly model did not show any better agreement with results from systematically collected data compared to the logistic regressions that used the same presence-inferred-absence data as the PresAbs-PresOnly.

Quality of inferred absences

Using absences which are in fact presences (false absences) can negatively affect model performance (Lobo, Jiménez-Valverde & Hortal 2010). We aimed to minimize false absences by using observations from reporters stating that they 1) consistently reported Siberian jay and, 2) were skilled in identifying it. Of the reporters that answered our questionnaire, 16% stated that they did not consistently report Siberian jay. Questionnaire recipients were selected because of their comparatively high reporting contribution and it is likely that the percentage that do not consistently report Siberian jay is higher amongst all reporters. We also expect that the listing of Siberian jay in both the 2005 and 2010 national red lists of Swedish species (Gärdenfors 2005; Gärdenfors 2010) has positively influenced the willingness to report the species. Consistent reporting rates may therefore be lower for many other species.

Highly skilled reporters have higher detection rates (Johnston *et al.* 2017), which also minimizes false absences. The high data contributions (> 2 million records during the study period) of our 38 reporters and the fact that several of them also took part in the SBS suggests that they are highly skilled in bird identification. This further suggests that inferred absences from their observations were of high quality. We thus recommend taking reporting consistency and species identification skills of reporters into account in order to minimize false absences when obtaining inferred absences.

For presence records of Siberian jay, the effect of variable reporter skills may be low as the species is comparatively easy to identify. For another relatively easy-to-identify species, the fox squirrel (*Sciurus niger*) in Florida, presence-only data from amateurs was as reliable as those from professionals (Tye *et al.* 2017).

Siberian jay conservation

An important conservation message many of our models suggest is that maintaining old forest patches within areas larger than individual home ranges is important to facilitate continued occupancy of suitable habitat. The positive effect of forest age in the mature forest patches surrounding an observation unit (variable Neighbourhood with 10 km resolution) was due to ages much larger than our threshold value (50 years). Habitat suitability for Siberian jay also increased with elevation, which we believe is a proxy for the percentage of mature forest in the larger landscape. The explanation might be a population level effect where larger areas of suitable habitat have an additional effect, for example by increasing reproductive success.

Availability of inferred absences for other species

CS datasets tend to be dominated by contributions from few very active reporters (Isaac & Pocock 2015). In our data, the 38 reporters providing inferred absences also provided about one third of all available Siberian jay presences. Keen reporters are frequently highly skilled in species identification and motivated to consistently report the less common species. Therefore, our results are likely relevant for species other than Siberian jay or birds, as sufficient inferred absences can likely be produced for many less common species across many taxa. This has been demonstrated for a species in a group that is less popular with reporters, a fungus (Mair *et al.* 2017b).

We found that not only the number of inferred absences, but also other properties, most likely their location (in environmental space) influenced results. This suggests that inferred absences in sparsely sampled regions are disproportionally important. Encouraging reporters to report from areas "off the beaten reporting track" could therefore likely provide large benefits for habitat suitability models.

Relevance for global presence-only data

Systematically collected data are not available for many regions and species. Encouraging keen reporters to consistently report the less common species, recruiting keen reporters in under-sampled regions and taking reporting consistency and species identification skills of reporters into account may be a suitable alternative for the modelling of the distributions of many species worldwide. The availability of inferred absences is not dependent on the existence of a checklist, but on consistent reporting of individual species by keen reporters. Importantly, inferred absences can therefore be obtained for species in less well-studied regions where even highly skilled reporters may not be able to identify all species of a group or where the knowledge about the species occurring is incomplete.

Authors' contributions

TS, UB, LM and MJ designed the study. UB analysed the data and wrote the first draft. All authors advised on models and contributed to the manuscript.

Acknowledgements

We thank the reporters of the Swedish Species Observation System (SSOB), the participants of our questionnaire and the volunteers and coordinators of the SBS (Lund University). Johan Nilsson extracted some observations from SSOS. Johan Nilsson, Johan Södercrantz and Ragnar Hall improved the questionnaire. We thank Michael Griesser and Nick Isaac for interesting discussions and two anonymous reviewers for constructive comments. Larger GIS calculations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). The study was funded by FORMAS grant 2012-991, and 2013-1096 to TS.

Data accessibility

The species data is deposited in the Dryad Digital Repository https://doi.org/10.5061/dryad.4722kf7. The forest data can be downloaded from ftp://salix.slu.se/download/skogskarta; data on urban areas from https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/geodata/oppna-geodata/tatorter/, data on settlements from https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/geodata/oppna-geodata/smaorter/, data on population density from https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/geodata/oppna-geodata/total-befolkning-per-ruta/; data on roads from http://www.lantmateriet.se/en/Maps-and-geographic-information/Maps/oppna-data/hamta-oppna-geodata/#faq:gsd-vagkartan-vektor; data on elevation from http://www.lantmateriet.se/en/Maps-and-geographic-information/Maps/oppna-data/hamta-

oppna-geodata/#faq:gsd-hojddata-grid-50; and climate data is available from https://esg-dn1.nsc.liu.se/projects/esgf-liu/ (search for 'mesan' after selecting a Federated ESFG-CoG Node).

References

Amano, T., Lamming, J.D.L. & Sutherland, W.J. (2016) Spatial gaps in global biodiversity information and the role of Citizen Science. *Bioscience,* **66,** 393-400.

Bibby, C.J., Burgess, N.D. & Hill, D.A. (1992) *Bird Census Techniques*. Academic Press.

Bird Life International (2016) *Perisoreus infaustus*. *The IUCN Red List of Threatened Species 2016***,** www.iucnredlist.org. Accessed 22 Sep 17.

Braunisch, V., Home, R., Pellet, J. & Arlettaz, R. (2012) Conservation science relevant to action: a research agenda identified and prioritized by practicioners. *Biological Conservation,* **153,** 201-210.

Brotons, L., Mönkkönen, M., Huhta, E., Nikula, A. & Rajasärkkä, A. (2003) Effects of landscape structure and forest reserve location on old-growth forest bird species in Northern Finland. *Landscape Ecology,* **18,** 377-393.

Burnham, K.P. & Anderson, D.R. (2004) Multimodel inference : Understanding AIC and BIC in model selection. *Sociological Methods & Research,* **33,** 261-304.

Cramp, S. & Perrins, C.M. (1994) *The Birds of the Western Palearctic*. Oxford University Press, Oxford.

Devictor, V., Whittaker, R.J. & Beltrame, C. (2010) Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions,* **16,** 354-362.

Edenius, L., Brodin, T. & White, N. (2004) Occurrence of Siberian jay *Perisoreus infaustus* in relation to amount of old forest at landscape and home range scales. *Ecological Bulletins,* **51,** 241-247.

Elith, J. & Leathwick, J. (2009) The contribution of species distribution modelling to conservation prioritization. In A. Moilanen, K.A. Wilson & H.P. Possingham (Eds.), *Spatial conservation prioritization* (pp. 70-93). Oxford University Press, New York.

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. & Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologistss. *Diversity and Distributions,* **17,** 43-57.

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution,* **6,** 424-438.

Fletcher Jr, R.J., McCleery, R.A., Greene, D.U. & Tye, C.A. (2016) Integrated models that unite local and regional data reveal larger-scale envrionmental relationships and improve predictions of species distributions. *Landscape Ecology,* **31,** 1369-1382.

Franklin, J. (2009) *Mapping species distributions - Spatial inference and prediction*. Cambridge University Press, Cambridge.

Gärdenfors, U. (2005) *The 2005 Red List of Swedish Species*. Swedish Species Information Centre, Uppsala.

Gärdenfors, U. (2010) *The 2010 Red List of Swedish Species*. Swedish Species Information Centre, Uppsala.

Griesser, M. & Lagerberg, S. (2012) Long-term effects of forest management on territory occupancy and breeding success of an open-nesting boreal species, the Siberian jay. *Forest Ecology and Management,* **271,** 58-64.

Griesser, M., Nystrand, M., Eggers, U. & Ekman, J. (2007) Impact of forestry practices on fitness correlates and population productivity in an open-nesting bird species. *Conservation Biology,* **21,** 767-774.

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling,* **135,** 147-186.

Higa, M., Yamaura, Y., Koizumi, I., Yabuhara, Y., Senzaki, M. & Ono, S. (2015) Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. *Diversity and Distributions,* **21,** 46-54.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2016) dismo: Species Distribution Modeling. R package version 1.1-1. http://CRAN.R-project.org/package=dismo.

Isaac, N.J.B. & Pocock, M.J. (2015) Bias and information in biological records. *Biological Journal of the Linnean Society,* **115,** 522-531.

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P. & Roy, D.B. (2014) Statistics from citizen science: extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution,* **5,** 1052-1060.

Johnston, A., Fink, D., Hochachka, W.M. & Kelling, S. (2017) Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution,* **00,** 1-10.

Kadoya, T., Ishii, H.S., Kikuchi, R., Suda, S.-i. & Washitani, I. (2009) Using monitoring data gathered by volunteers to predict the potential distribution of the invasive alien bumblebee *Bombus terrestris*. *Biological Conservation,* **142,** 1011-1017.

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T. & Donald, P.F. (2016) Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Diversity and Distributions,* **22,** 1024-1035.

Kéry, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography,* **2010,** 1851-1862.

Kéry, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Häfliger, G. & Zbinden, N. (2010) Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. *Conservation Biology,* **24,** 1388-1397.

Kéry, M. & Schaub, M. (2012) *Bayesian population analysis using WinBUGS*. Academic Press, Oxford, UK.

Lobo, J.M., Jiménez-Valverde, A. & Hortal, J. (2010) The uncertain nature of absences and their importance in species distribution modelling. *Ecography,* **33,** 103-114.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology,* **84,** 2200-2207.

Mair, L., Harrison, P.J., Bärring, L., Strandberg, G. & Snäll, T. (2017a) Forest management could conteract distribution retractions forced by climate change. *Ecological Applications,* **27,** 1485-1497.

Mair, L., Harrison, P.J., Jönsson, M., Löbel, S., Nordén, J., Siitonen, J., . . . Snäll, T. (2017b) Evaluating citizen science data for forecasting species responses to national forest management. *Ecology and Evolution,* **7,** 368-378.

Mair, L. & Ruete, A. (2016) Explaining spatial variation in the recording effort of Citizen Science Data across multiple taxa. *PLOS One,* **11 (1),** e0147796.

Nystrand, M., Griesser, M., Eggers, U. & Ekman, J. (2010) Habitat-specific demography and source-sink dynamics in a population of Siberian jays. *Journal of Animal Ecology,* **79,** 266-274.

Ottvall, R., Green, M.O., Lindström, Å., Esseen, P.-A. & Marklund, L. (2007) Landskapets betydelse för fåglarnas förekomst och populationsutveckling: en pilotstudie med monitoringdata från Svensk Fågeltaxering och NILS. Rapport. Ekologiska institutionen, Lunds universitet. 53 pages.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling,* **190,** 231-259.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography,* **31,** 161-175.

Phillips, S.J., Dudik, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications,* **19,** 181-197.

Plummer, M. (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)***,** March 20-22, Vienna.

Polce, C., Termansen, M., Aguirre-Gutiérrez, J., Boatman, N.D., Bugde, G.E., Crowe, A., . . . Biesmeijer, J.C. (2013) Species distribution models for crop pollination: a modelling framework applied to Great Britain. *PLOS One,* **8(10),** e76308.

Pukkala, T., Sulkava, R., Jaakkola, L. & Lähde, E. (2012) Relationships between economic profitability and habitat quality of Siberian jay in uneven-aged Norway spruce forest. *Forest Ecology and Management,* **276,** 224-230.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmermann, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography,* **33,** 1689-1703.

Reese, H., Nilsson, M., Pahlén, T.G., Hagner, O., Joyce, S., Tingelöf, U., . . . Oslsson, H. (2003) Countrywide estimates of forest variables using satellite data and field data from the National Forest Inventory. *Ambio,* **32,** 542-548.

Silvertown, J. (2009) A new dawn for citizen science. *Trends in Ecology and Evolution,* **24,** 467-471.

Snäll, T., Forslund, P., Jeppsson, T., Lindhe, A. & O'Hara, R.B. (2014) Evaluating temporal variation in Citizen Science Data against temporal variation in the environment. *Ecography,* **37,** 293-300.

Snäll, T., Kindvall, O., Nilsson, J. & Pärt, T. (2011) Evaluating citizen-based presence data for bird monitoring. *Biological Conservation,* **144,** 804-810.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D. & Kelling, S. (2009) eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation,* **142,** 2282-2292.

Syfert, M.M., Smith, M.J. & Coomes, D.A. (2013) The effects of sampling bias and model complexity on the predictive performance of Maxent species distribution models. *PLOS One,* **8,** e55158.

Tye, C.A., McCleery, R.A., Fletcher Jr, R.J., Greene, D.U. & Butryn, R.S. (2017) Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology,* **54,** 628-637.

van Strien, A.J., van Swaay, C.A.M. & Termaat, T. (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology,* **50,** 1450-1458.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Grant, E.H.C. & Veran, S. (2013) Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution,* **4,** 236-243.

Yoccoz, N.G., Nichols, J.D. & Boulinier, T. (2001) Monitoring of biological diversity in space and time. *Trends in Ecology & Evolution,* **16,** 446-453.

Zadrozny, B. (2004) Learning and evaluating classifiers under sample selection bias. *Proceedings of the 21st International Conference on Machine Learning***,** 114.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. & Smith, G.M. (2009) *Mixed Effects Models and Extensions in Ecology with R*. Springer.

Table 1:

Environmental predictor variables in models for Siberian jay occurrence. All variables were calculated for observation units of 2x2 km with the exception of Neighbourhood, which was calculated in moving windows of 10x10 km. For details of calculations and data sources, see Appendix S4.

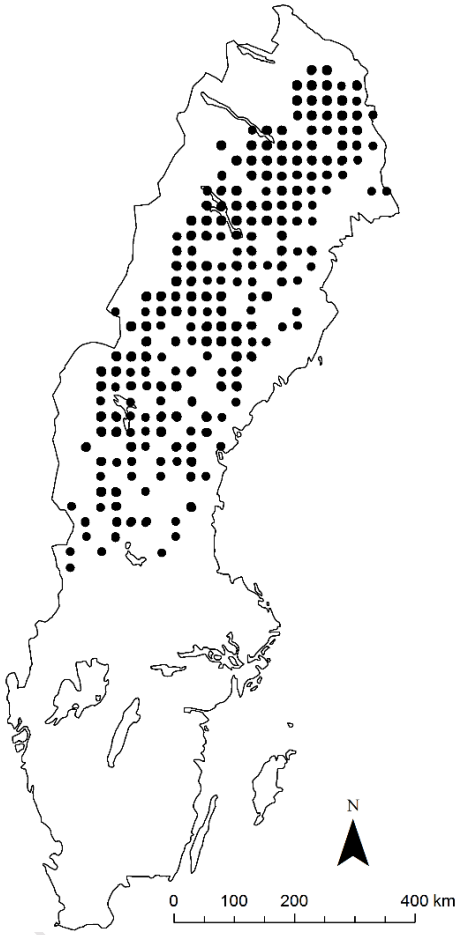
Category	Description	Abbreviation
Mature forest	Percentage mature forest (>=50 years & >=100 m^3/ha)	PercMature
	Mean forest age	MeanAge
	Mean forest volume	MeanVol
	Mean age of patches (>=50 years & >=100 m^3/ha & >30 ha)	Neighbourhood
Forest	Percentage non-mature forest (< 50 years or < 100 m^3/ha)	PercOther
	Percentage spruce volume on total volume	PercSpruce
	Percentage pine volume on total volume	PercPine
Non-forest	Winter temperature (January + February)	WinterTemp
	Spring precipitation (April + May)	SpringPrec
	Distance to nearest settlement	DistSettl
	Elevation	Elevation

Table 2:

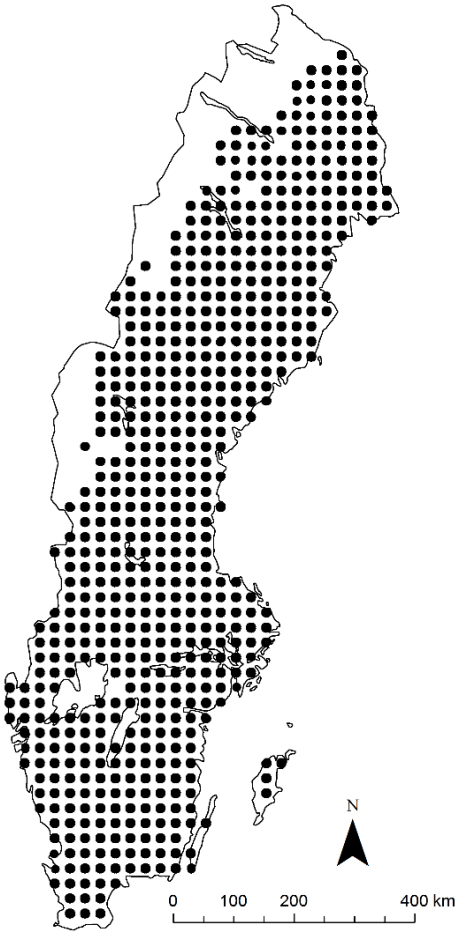
Regression coefficients or lambda values (MaxEnt) of standardized variables in the best model per Siberian jay modelling method. Association with directions contrary to expectations are highlighted in red. The PresAbs-PresOnly model includes the bias variables PopDen and DistRoad. The site-occupancy-detection model includes log(NoVisits) as a variable for the detection process. For variable definitions, see Table1 and Table S2 (for bias variables).

		SBS	PresAbs-38	PresAbs-all	PresAbs-PresOnly	MaxEnt-TGB	MaxEnt-Ran	Occ-Det
	Intercept	-3.83	-2.30	-0.80	-2.37			-3.01
Mature Forest	PercMature	0.28	0.61	0.28	0.19	1.32	0.60	0.49
	PercMature^2		-0.10		-0.06	-0.68	-1.73	0.31
	MeanAge	0.44			-0.15	-0.71		
	MeanVol			0.48			2.15	0.89
	MeanVol^2							
	Neighbourhood		0.12	0.18	0.01		0.54	-0.74
Forest	PercOther	0.27	0.17	0.18	-0.23	-0.38	-0.54	0.84
	PercSpruce				0.56			-0.33
	PercSpruce^2				-0.28			
	PercPine			0.09	0.25			
	PercPine^2			0.08	0.18			
Non-forest	WinterTemp	-1.65	-2.30	-2.04	-1.73	-4.82	-4.81	-3.33
	SpringPrec	-0.51		0.01	-0.07	-1.02	-0.96	
	WinterTemp * SpringPrec	-0.49		-0.21	-0.09	-2.63	-2.19	
	DistSettl		-0.15	-0.25	-0.08	-0.59	-2.08	
	Elevation	1.55	1.58	1.76	0.90	5.19	5.70	2.38
	Elevation^2	-0.46	-0.39	-0.28	-0.21	-4.03	-8.48	

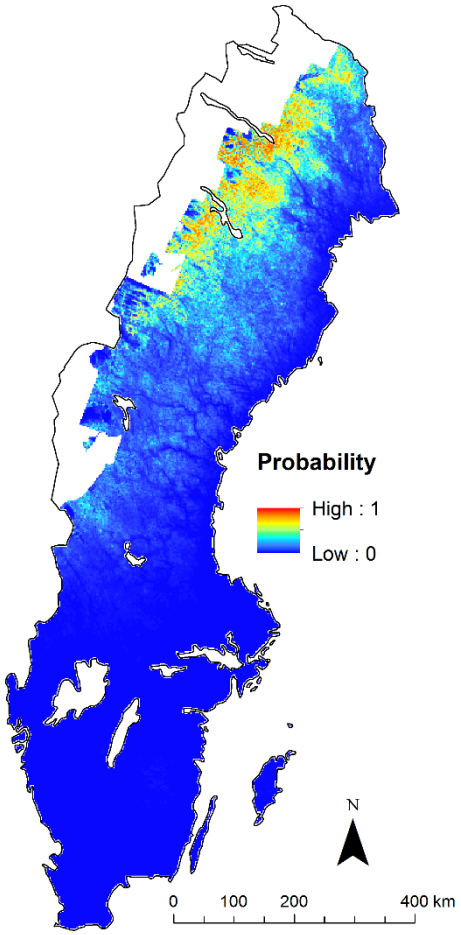
a) SBS presences



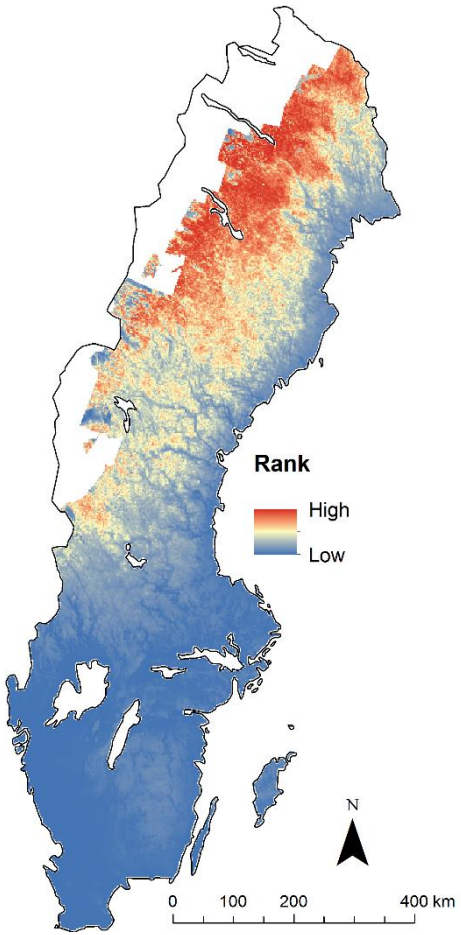
b) SBS absences



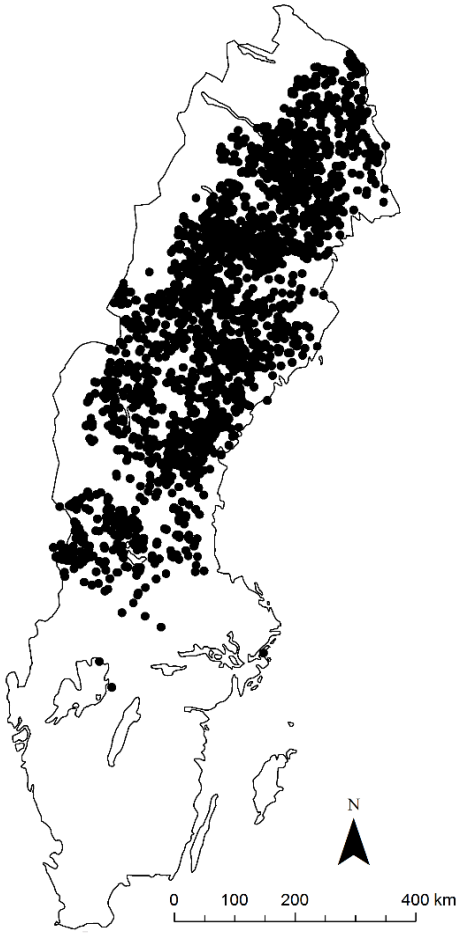
c) SBS distribution



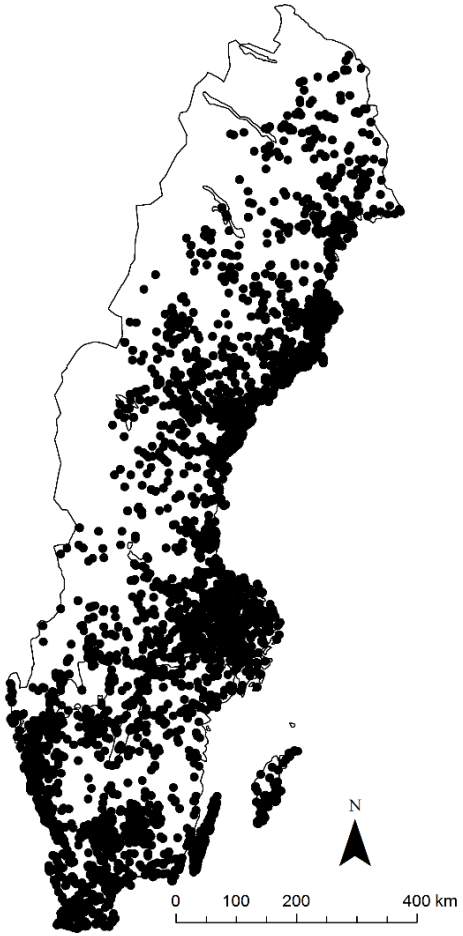
d) SBS habitat suitability



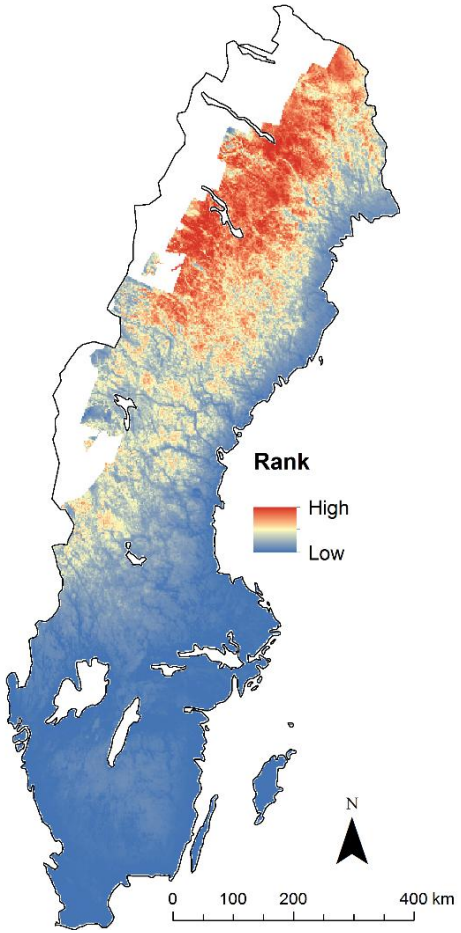
e) CS presences



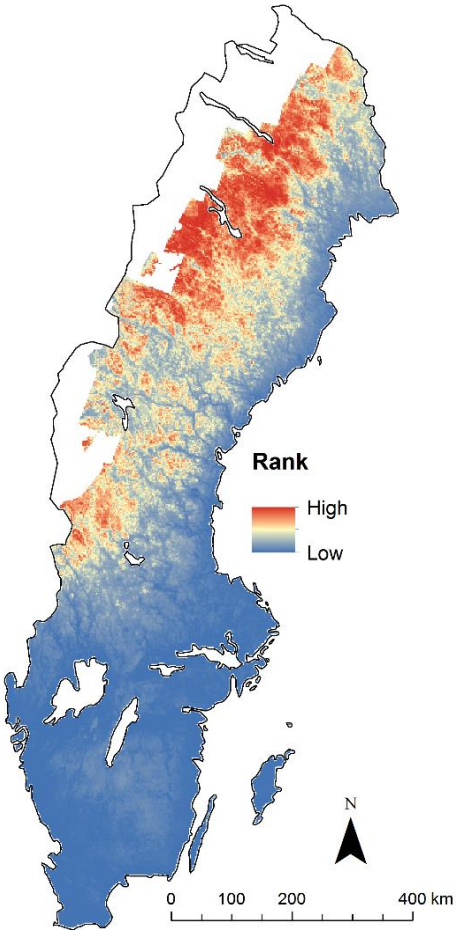
f) CS inferred absences



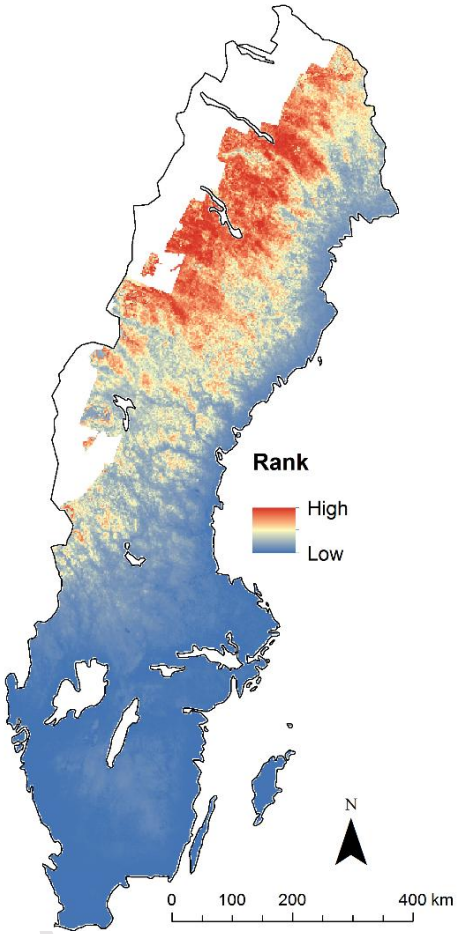
g) PresAbs-38 habitat suitability



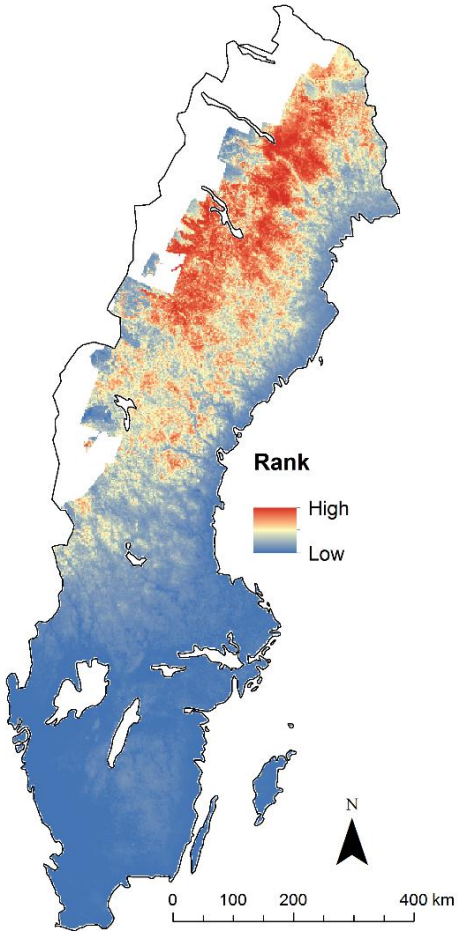
h) PresAbs-all habitat suitability



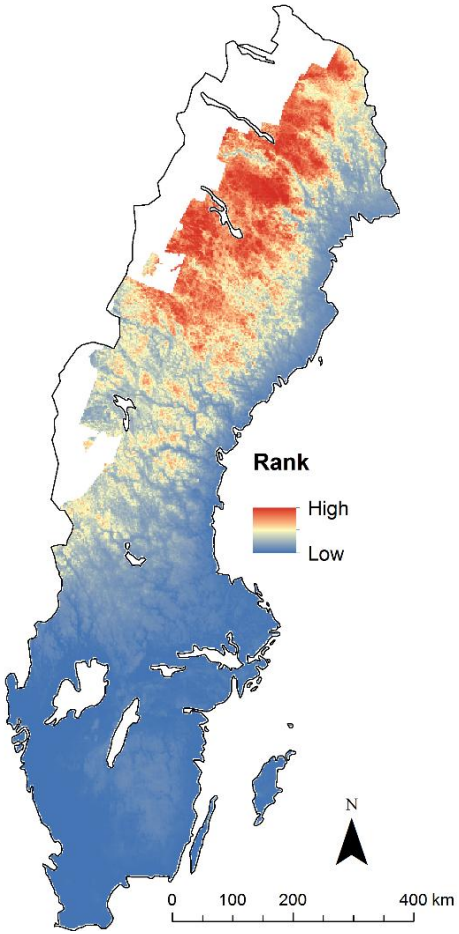
i) PresAbs-PresOnly habitat suitability



j) MaxEnt-Random habitat suitability



k) MaxEnt-TGB habitat suitability



l) Occ-Det habitat suitability

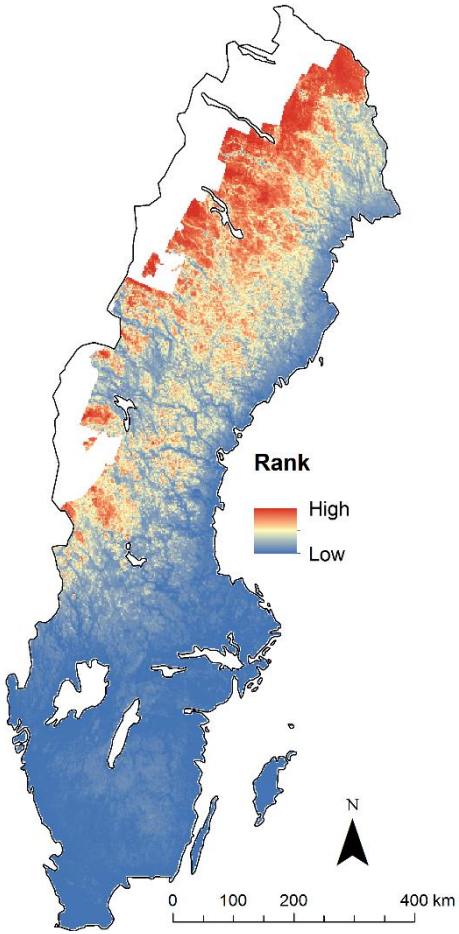
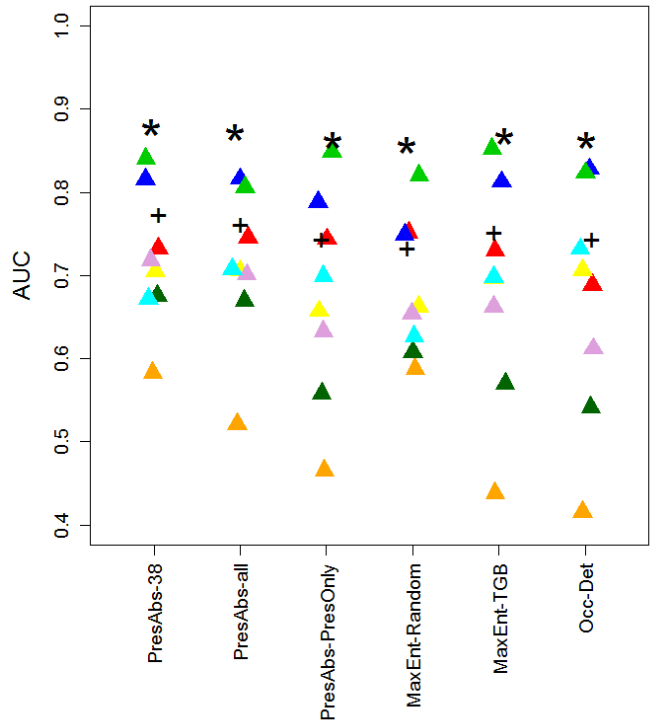
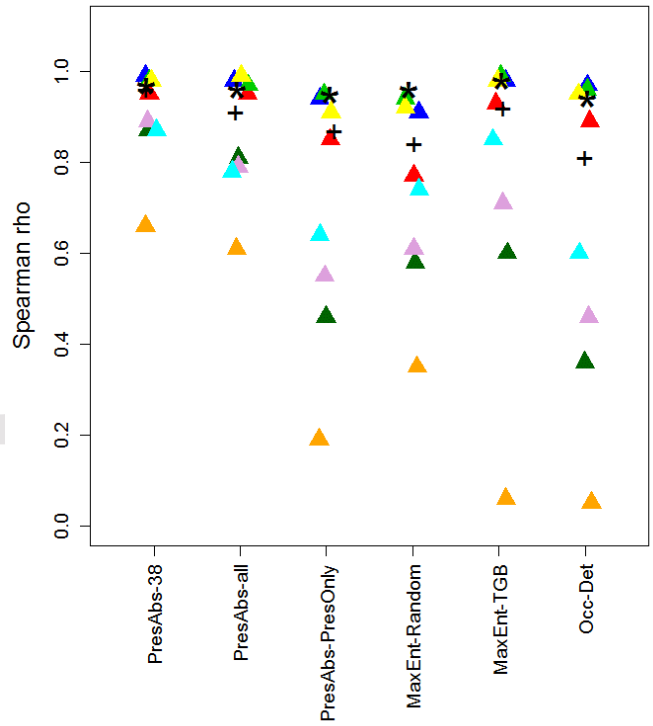


Fig 1: a) Presence and b) absence data from the SBS, predicted Siberian jay distribution using the SBS model as c) probability and d) habitat suitability. e) Presence and f) inferred absence data from CS and g - l) predicted habitat suitability using the CS models. Map resolution: 2x2 km. Habitat suitability scores for each grid cell were converted to ranks to facilitate visual comparison between habitat suitability maps and are presented with the same colour scheme and gamma stretch.

a)



b)



c)

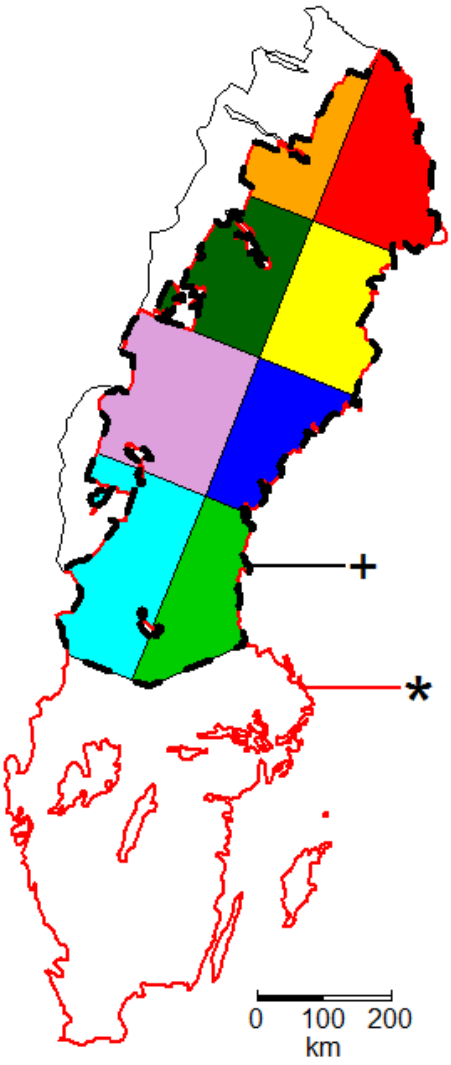


Fig. 2: a) The agreement of habitat suitability scores per CS method with systematically collected SBS observations using AUC and b) the agreement between habitat suitability maps per CS method with a habitat suitability map from SBS observations using Spearman rank correlation coefficient. c) Geographic areas to which colours and symbols in a) and b) correspond: *: study area (red line), +: Siberian jay range within the study area (dashed black line).